Larger languages have higher entropy.

Gary Lupyan
Department of Psychology
University of Wisconsin-Madison

The transmission of language is constrained by what people can learn. While all languages are constrained by the learning biases of young children, languages that have adult learners are additionally constrained by what can be learned by nonnative speakers. It has been argued that the languages spoken by more people are morphologically simpler because of a selection against morphological complexity, which is difficult for adults to learn (e.g., Lupyan & Dale, 2010; Trudgill, 2001). But this does not explain why so many languages are so complex to begin with. Here, we build on our previous hypothesis that morphological complexity is a form of redundancy that may facilitate language learning in young children (Dale & Lupyan, 2012; Lupyan & Dale, 2010). On this account, the greater complexity observed in smaller languages that are not constrained by what adults can learn, may be a functional adaptation to increase learnability by young children by decreasing entropy thereby lowering processing costs. We measured entropy by compressing written language corpora using the LZW algorithm (see Ehret & Szmrecsanyi, 2016; Juola, 1998 for similar approaches). Lower entropy is marked by higher compressibility: a highly compressible language is one in which one can predict the occurrence of current elements from past elements, and thus the current elements are contributing less information. We measured the compressibility of 1268 languages using the Parallel Bible Corpus (Mayer & Cysouw, http://paralleltext.info/). We find that languages differ substantially in their compressibility (Fig 1A) and that there is a strong relationship between compressibility and population size (Fig. 1B). This relationship survives controls for differences in orthography, vocabulary size, and areal, and phylogenetic confounds (adjust b = .017, t>6). To our surprise, redundancy was largely unrelated to morphological complexity. Rather than arising from morphology, the lower entropy (greater redundancy) of languages with few speakers was a function of these languages having more frequent and longer repeated strings. Testing the hypothesis that this type of redundancy aids language learning or processing in young children requires further empirical work. However, our results provide strong evidence that population correlates not only with grammatical structures, but also with information-theoretic properties of language.

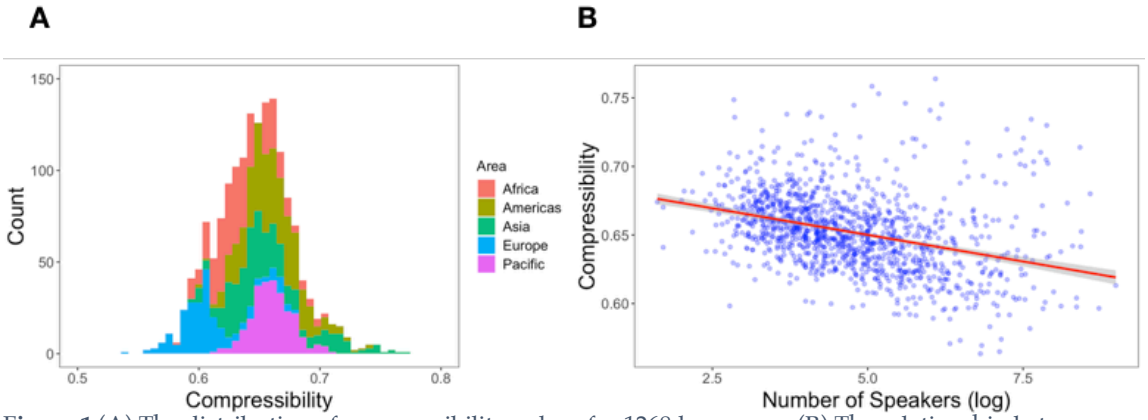

Figure 1 (A) The distribution of compressibility values for 1268 languages. (B) The relationship between population size and compressibility (larger compressibility = greater redundancy = lower entropy).